Methoden für Trendanalysen im Web zur Unterstützung des Customer Relationship Management

Kai Heinrich
Andreas Hilbert
Martin Kersten

Veröffentlicht in:
Multikonferenz Wirtschaftsinformatik 2012
Tagungsband der MKWI 2012
Hrsg.: Dirk Christian Mattfeld; Susanne Robra-Bissantz

Braunschweig: Institut für Wirtschaftsinformatik, 2012

Methoden für Trendanalysen im Web zur Unterstützung des Customer Relationship Management

Kai Heinrich

TU-Dresden, Fakultät Wirtschaftswissenschaften, 01187 Dresden,
E-Mail: kai.heinrich@tu-dresden.de

Andreas Hilbert

TU-Dresden, Fakultät Wirtschaftswissenschaften, 01187 Dresden,
E-Mail: andreas.hilbert@tu-dresden.de

Martin Kersten

TU-Dresden, Fakultät Wirtschaftswissenschaften, 01187 Dresden,
E-Mail: kai.heinrich@tu-dresden.de

Abstract

Mit dem Einzug des Web 2.0 ins tägliche Leben haben Individuen die Möglichkeit ihre Meinungen und Gefühle in Form von Blogs zu veröffentlichen. Die Analyse der Trends in dieser Blogosphäre kann maßgeblich zur Unterstützung der Kundenrückgewinnung in einem CRM-System eingesetzt werden. In dieser Forschungsarbeit werden bestehende Ansätze zur Trenderkennung im Allgemeinen untersucht und anschließend die Eignung ihrer Applikation auf Weblogs geprüft. Dazu wird ausgehend von bestehenden wissenschaftlichen Arbeiten ein System zur Trendanalyse prototypisch implementiert und die Analyseergebnisse im Anschluss evaluiert.

1 Einleitung

1.1 Motivation

Das explosive Wachstum des Internets und die große Verbreitung von Social Media Lösungen haben neue Möglichkeiten für Menschen geschaffen, ihre Meinungen online zu verbreiten (vgl. [32]). Weblogs, RSS, Screencasting oder Video-Blogging, Podcasting, SocialBookmarking, Tagging, all dies sind neue Technologien, die im World Wide Web zu finden sind (vgl. [1]). Diese Instrumente sind Werkzeuge geworden, welche die Art und Weise der Bereitstellung von Inhalten revolutioniert haben. Die Blogosphäre, welche die Gesamtheit der Blog-Websites verkörpert, ist

eine große Quelle für Trendanalysen in Bereichen wie Produkt-Umfrage, Kundenbeziehungsmanagement und Marketing geworden. Eine Besonderheit der Blogosphäre ist ihre große Dynamik gegenüber traditionellen Webseiten. Eine Ankündigung eines neuen Produktes oder ein Geschehen in der Welt können unter Umstanden sofortiger Auslöser von intensiven Diskussionen über verschiedenste Blogseiten sein. Derzeit existieren über 140 Millionen Blogs und rund 100 Tausend neue Blogs kommen jeden Tag neu dazu. Die Beobachtungen der letzten Jahre haben gezeigt, dass sich die Anzahl der vorhandenen Blog-Seiten alle 200 Tage verdoppelt hat (vgl. [32]). In diesen Blogs schreiben Menschen über diverse Themen wie ihr persönliches Leben, Testberichte zu verschiedenen Produkten und Meinungen über Politik, Sport oder Lifestyle. Überwachung und Analyse von Informationen innerhalb dieser Datenbestände können wichtige Informationen und wertvolle Erkenntnisse zur öffentlichen Meinung liefern und damit das Customer Relationship Management (CRM) unterstützen (vgl. [27]).

1.2 Forschungsdesign

Es existiert ein Defizit von tauglichen Verfahren zur Analyse von Texten die Trends und Veränderungen im Verlauf erkennen können. Daraus resultiert das Gestaltungsproblem, ein System zur Bewältigung dieser Aufgabe zu entwickeln. Für diese Entwicklung ist es notwendig, vorhandene Ansätze zu analysieren und hinsichtlich ihrer Funktionalität zu evaluieren. Aus diesen beiden die Untersuchungen bestimmenden Problemen resultieren folgende Forschungsfragen:

- Welche Eigenschaften kennzeichnen einen Trend?
- Welche Methoden bzw. Algorithmen sind zur Identifikation von Trends und Trendbrüchen in Texten geeignet?
- Wie aussagekräftig sind diese Methoden?

Zur Beantwortung dieser Fragen werden zunächst die Anatomien von Trends sowie Methoden zur Analyse von Trends vorgestellt. Diese Methoden werden anschließend mit Hilfe der Implementierung des Systems evaluiert.

2 Grundlagen zur Trendbetrachtung

2.1 Anatomie eines Trends

Ein Trend wird prinzipiell über eine Bezugsgröße dargestellt. Beispielsweise wird der Trend „steigendes Wirtschaftswachstum“ durch die Bezugsgröße „Bruttoinlandprodukt“ dargestellt. Jede Veränderung der Bezugsgröße basiert auf Ursachen, welche sich im Trend wiederspiegeln. Je besser man die Ursachen und Hintergründe kennt und versteht, desto bessere Aussagen lassen sich für die Zukunft prognostizieren. Um das zu erreichen, müssen die Treiber (Was treibt den Trend an) der Veränderung identifiziert werden (vgl. [25]).

Die Ursachen der Veränderung zu erkennen und deren Wirkung zu verstehen, ist jedoch nur ein Teilprozess in der Diagnose des Trendermittlungsprozesses.

Die „Treiber“ bestimmen bzw. definieren den Trend und dessen Verlauf. Sobald diese Wirkungskräfte ermittelt wurden, beginnt der Prozess der Hypothesenbildung (vgl. [25]). Dieser Prozess teilt sich in zwei temporale Kategorien. Der erste ist das Verständnis der Vergangenheit bzw. der vollständigen Diagnose der vergangenen Merkmalsausprägungen.

In diesem beschriebenen Prozess ist die Trendidentifikation in Texten eindeutig ein Problembereich der Diagnose. Ziel der Applikation soll es sein, die auftretenden Merkmalsausprägungen zu analysieren, um eine „Beschreibung“ der Veränderung zu ermöglichen, potentielle „Treiber“ zu identifizieren und zu entscheiden, ob es sich letzten Endes bis zum derzeitigen Beobachtungszeitraum um einen Trend handelt oder nicht. Der Blick in die mögliche Zukunft des Trends beginnt erst mit der Projektion der Daten, in der kurzfristige oder langfristige Prognosen gebildet werden. Diese Prognosen bilden somit den zweiten Teil des oben genannten Prozesses. Für das CRM ist diese Analyse ein wichtiger Schritt im Rahmen der Kundenrückgewinnung, da in der Regel nicht feststeht, wann ein Kritikpunkt zur Abwanderung eines Kunden führt (vgl. [29]). Das frühzeitige Erkennen von möglichen Ursachen durch Analyse der Inhalte in Web-Daten könnte aber wichtige Erkenntnisse über potentielle Abwanderungsursachen liefern.

2.2 Trendbruch

Anders als bei Trends, die eine kontinuierliche Entwicklung darstellen, sind Trendbrüche radikale oder abrupte Änderungen im Verlaufsmuster (vgl. [25]). [24] beschreibt diese strukturellen Veränderungen als Wendepunkte, die sich durch eine Veränderung der Bezugsgröße durch sich selbst, ihrer eigenen Vergangenheit oder ihrer relativen Position auf dem Zeitpfad erklärt (vgl. [24]). Diese Diskontinuitäten im Trend können den weiteren Verlauf des Trends maßgeblich beeinflussen und so eine klare Hypothesenbildung verfälschen, wenn diese nicht berücksichtig werden (vgl. [5]). Zum einen wird der Trend nach dem Trendbruch (Trendfolgeformation) weiter fortgesetzt und zum anderen kann der Trendbruch ein Indiz für eine Umkehr des Trends sein (Trendumkehrformation). Bei der Trendfolgeformation handelt es folglich um eine zeitliche bzw. vorübergehende Trendunterbrechung, im Gegensatz zur Trendumkehrformation, in der ein Strukturbruch im Trend zu einem Trendwechsel mutiert (vgl. [6]). Grundsätzlich sollte im Kontext der Analyse von Meinungen in Web-Daten jeder größere Impuls oder Strukturbruch (Signale) in den Merkmalsausprägungen des Trendverlaufs auf Ursachen analysiert werden, um ggf. einen Trendwechsel ins Negative verhindern oder ins Positive fördern zu können. Die Beschreibungsinstrumente der Zeitreihenformationen unterstützen dabei das bessere Verstehen und die Interpretation der Veränderungen.

2.3 Konzept von Ansoff

Der strategische Nutzen für das Unternehmen in Bezug auf die Unterstützung des CRM, der durch das Beobachten oder „Monitoren“ von Blogs entstehen kann, wird in der Literatur stark diskutiert. In Bezug auf das Verfahren der Interpretation von Informationen aus Blogs oder anderen Web-Quellen für Entscheidungen im Unternehmen, wird in der wissenschaftlichen Literatur das Konzept von ANSOFF häufig als eine Möglichkeit beschrieben (vgl. [1]). Dieses Konzept basiert auf der Annahme, dass strategische Entscheidungen nicht ausschließlich auf extrapolierten Trendbeobachtungen basieren sollten, sondern Diskontinuitäten, also Strukturbrüche in Trends, mit in die Entscheidungen einfließen sollten. Diskontinuitäten kündigen sich durch eine Veränderung der Umwelt (schwache Signale) an, bei denen die Bedeutung und die Auswirkung ungewiss sind. Weil sich diese Signale mit der Zeit verdichten und konkreter werden, lohnt sich eine frühzeitige Erfassung dieser schwachen Signale, auch wenn sie zu dem Zeitpunkt noch nicht Interpretierbar sind (vgl. [30]). Aus diesem Grund schlägt [1] ein System der Frühaufklärung und -erkennung vor (Weak Signal Management). Dieses Konzept hat zum Ziel, durch Abtasten der Umwelt, was als sogenanntes „Scanning“ bezeichnet wird, diese schwachen Signale zu erkennen und zu verarbeiten (vgl. [9]). Im Kontext der Analyse von Blogs entstehen

solche schwachen Signale in den Veränderungen von Meinungen in Diskussionen (vgl. [23]). Wenn sich die schwachen Signale soweit verdichten lassen, dass sich ein „Strategic Issue“ erkennen lässt, wird es Zeit, dieses Thema weiterzuverfolgen und Wirkungen bzw. Konsequenzen zu analysieren (vgl. [30]).

3 Forschungsansätze zur Trendanalyse

3.1 Burst-Analyse

Die Verbreitung von Informationen im Web läuft aufgrund der Vielzahl von Autoren unkoordiniert ab. Dennoch gilt das allgemeine Phänomen, dass wenn Ereignisse von großem Interesse stattfinden, verstärkt darüber berichtet wird und somit der Informationsgehalt über dieses Thema ansteigt (vgl. [22]). Daraus resultiert, dass die Beliebtheit von gewissen Schlüsselwörtern größer wird. Dieser Anstieg der vorkommenden Häufigkeit von Schlüsselwörtern über einen begrenzten Zeitraum wird „Burst" bzw. „Ausbruch“ genannt (vgl. [7]). Ziel der Burst-Analyse ist es, diese Ausbrüche und das Ereignis, das damit zusammenhängt, zu identifizieren (vgl. [3]). [26] beschreiben die Burst-Analyse als einen Prozess der Identifizierung von kurzeitig auftretenden Begriffsanomalien auf Basis von Langzeitbeobachtungen (vgl. [26]). Das Aufdecken solch einer unerwarteten Beliebtheit von bestimmten Begriffen im Web kann von großer Bedeutung sein, weil diese Analyse Strukturbrüche im Trend identifizieren und Indikatoren für sich entwickelnde Trends liefern kann. [3] unterscheiden dabei zwei Arten von Ausbrüchen: vorweggenommene und überraschende Ausbrüche. Bei vorweggenommen Ausbrüchen wird die auftretende Häufigkeit ständig größer, erreicht nach gewisser Zeit ein Maximum und sinkt anschließend wieder. Beispiele dafür sind u.a. die Veröffentlichung von einem Film oder die Einführung eines lang angekündigten Produkts von gewisser Popularität. Überraschende „Bursts" sind nicht vorhersehbar und treten völlig unerwartet auf. Auf Basis des Ansatzes von [15] und dessen Weiterentwicklung durch [3] wird ein Verfahren vorgestellt, um „Bursts" zu identifizieren (vgl. [15] & [3]):

$$\mu = \frac{1}{w}\sum_{i=1}^{w} x_i \quad (1) \qquad \sigma^2 = \frac{1}{w}\sum_{i=1}^{w}(x_i - \mu)^2 \quad (2) \qquad \sigma = \sqrt{\sigma^2} \quad (3) \qquad f = \mu + 2\sigma \quad (4)$$

Es wird das arithmetische Mittel μ (vgl. (1)) über die absolut auftretenden Häufigkeiten eines Tages von einem Schlüsselwort x über einen Beobachtungszeitraum von w Tagen berechnet. Das Ergebnis μ sagt aus, wie oft der Begriff x im Durchschnitt pro Tag im Beobachtungszeitraum w vorgekommen ist. Im nächsten Schritt wird die mittlere quadratische Abweichung σ^2 (vgl. (2)) berechnet. Die Varianz ist ein Streuungsmaß für die Lage der Zeitreihenwerte x_i um den Durschnitt der Zeitreihe x_i. Um die Streuung verschiedener Zeitreihenwerte miteinander vergleichen zu können, ist es erforderlich, die Varianz auf ein einheitliches Maß, die Standardabweichung σ (vgl. (3)), zurückzuführen (vgl. [14]). Im dritten und letzten Berechnungsschritt wird mit Hilfe der in den vorherigen Schritten ermittelten Werte ein Faktor berechnet, mit dem die auftretenden Häufigkeiten verglichen werden, um festzustellen, ob dieser als Ausbruch klassifiziert wird oder nicht. Übersteigt ein Wert den Faktor f (vgl. (4)), so wird dieser als „Burst" kenntlich gemacht. Der Faktor unterliegt dem Gesetz der Standardnormalverteilung d.h. bei dem festgelegten Wert von 2σ liegt der Erwartungswert für zu erkennende Burst bei 2,3%.

3.2 Keyword-Analyse

Schlüsselwörter repräsentieren den Verlauf über ein Thema und können Schwerpunktinhalte zurückgeben (vgl. [4]). Der Zusammenhang von Schlüsselwörtern ist dabei nicht statisch. Je nach temporalem Beobachtungsabstand und dem Verlauf der Themenentwicklung ändert sich die Bedeutung der Schlüsselwörter. Bei der Keyword-Analyse ist es das Ziel, genau diese Zusammenhänge zu erkennen und wiederzugeben. Als Maß dafür, welche Bedeutung ein Schlüsselwort für ein Thema besitzt, wird die Korrelation herangezogen. Sind Korrelationen von Schlüsselwörtern zu einem Thema erkennbar, so spiegeln diese in ihnen auftretende Ereignisse wider (vgl. [4]). Im Kontext der Trendforschung können besonders häufig auftretende Schlüsselwörter ein Indiz für einen Trend oder Trendbruch sein. Gängige Verfahren, um statistische Zusammenhänge zwischen zwei Größen zu ermitteln wie z.B. Transinformation oder „mutual Information“ scheiden aufgrund der hohen Datenmenge mit Millionen von Dokumenten oder Textinhalte im Web aus. Term Frequency – Inverse Document Frequency (TF-IDF) ist ein Verfahren, das im „Information Retrieval“ zur Beurteilung der Bedeutung einzelner Terme in großen Textdaten angewendet wird (vgl. [28]). Im Rahmen der Keyword-Berechnung ist dieses Verfahren zur Term-Gewichtung die Grundlage (vgl. [21]). [3] haben diese Methode für die Berechnung von Korrelationseffekten so modifiziert, dass sich valide Schlüsselwörter über ein Suchwort bilden lassen, ohne den ganzen Datensatz überprüfen zu müssen (vgl. [3]):

$$G(W_q, W_n) = TF_{q,n} \cdot IDF_n \quad \text{mit} \quad TF_{q,n} = N_{DmW_q}(W_n) \quad \text{und} \quad IDF_n = \log \frac{|D|}{|D_{W_n}|} \quad (5)$$

$TF_{q,n}$ wird dabei durch die Frequenz des möglichen Keywords Wn in den Dokumenten der Datenbasis Dm, die das Suchwort Wq enthalten ausgedrückt. Anders ausgedrückt wird die Häufigkeit in den Dokumenten Dm ermittelt, in denen die beiden Wörter Wq und Wn gemeinsam vorkommen. Dm ist dabei eine Einschränkung der zu durchsuchenden Datenbasis. [3] verwenden für die Berechnung von potentiellen Schlüsselwörtern m=30 Dokumente. Dies bedeutet, dass nur die aktuellen m Dokumente oder Datensätze nach diesem Verfahren durchsucht werden, die ein Suchwort Wq enthalten. Im Kontext der Trendanalyse sollte jedoch überlegt werden, einen größeren Wert für m zu wählen bzw. die Einschränkung der Datenbasis nach bestimmten Beobachtungszeiträumen vorzunehmen, um Entwicklungen über die Dimension Zeit analysieren zu können. Die Terme mit den höchsten Gewichtungen werden als Schlüsselwörter oder Keywords für ein Suchwort identifiziert.

3.3 SVD-Analyse

Die Singulärwertzerlegung, auch Single Value Decomposition (SVD), ist ein optimales orthogonales Zerlegungsverfahren von Matrizen, um in großen Datenbeständen Informationen zu filtern und zu verdichten; dieses Verfahren wird häufig zur Problemlösung der Methode der kleinsten Quadrate auf große Datenmatrizen verwendet (vgl. [20]). Im Kontext der Identifikation von Trends und Trendbrüchen auf Basis von Webinhalten ist jedoch nur der Aspekt der Zeitreihenanalyse interessant. Der Schwerpunkt liegt dabei besonders in der Erkennung des Trends eines Schlüsselwortes in der Blogosphäre bzw. in Webdokumenten. Entwickelt wurde diese Abwandlung der Singulärwertzerlegung zur Trendanalyse von Schlüsselwörtern in verschiedenen Blogs von [10], basierend auf der Arbeit von [11], welche erstmalig die SVD nutzten, um große Textmengen zu analysieren.

3.4 Konzeption einer Applikation zur Trendanalyse

Der Analyseprozess lehnt sich dabei an das Vorgehen zur Entdeckung neuer Muster in Daten (KDD-Prozess) von [11] an. Zu Beginn wird die Zielstellung festgelegt, die mit Hilfe des Vorgehensmodells erreicht werden soll (vgl. [19]). Daraus resultiert je nach Anwendungsfall und Untersuchungsgegenstand die Auswahl der zugrundeliegenden Daten (vgl. [16]). Im Kontext der „Trendanalyse im Web zur Unterstützung des Customer Relationship Management“ wurden als Datengrundlage sogenannte Feeds benutzt. Diese Feeds ermöglichen es auf aktuelle Beiträge in Form von Weblogs in einheitlicher Form zuzugreifen. Bedingung war es, dass diese Feeds täglich neue Einträge enthalten, um eine kontinuierliche Datengewinnung gewährleisten zu können. Nachdem das angestrebte Ziel formuliert und die Datengrundlage bestimmt wurde, beginnt der eigentliche Prozess mit der Datenselektion. Hinsichtlich der Zielfragestellung und der Datengrundlage, welche aus Feeds besteht, wird in diesem Prozessabschnitt nur eine horizontale Datenselektion vorgenommen. Laut der Definition von Hagemann [17] ist die Selektion dann als optimal zu erachten, „wenn alle Attribute ausgewählt wurden, die für die zu beantwortenden Fragen nötig sind, und keine Attribute, die dafür überflüssig sind“ ([17]). Mit der Auswahl der vier benötigten „Tags“ oder Attribute Content (Inhalt), Titel, Datum und der Link (Feed-Quelle) ist in diesem Fall das Optimum gegeben. In dem gesamten Beobachtungszeitraum der Feed-Quellen wurden automatisiert über 6000 Aktualisierungen vorgenommen, in denen 531.948 unterschiedliche Einträge mit den „Tags“: Titel, Content, Datum und Link (Feed-Quelle) gespeichert wurden. Im Aufbereitungsschritt geht es in erster Linie darum, die Qualität des Datenbestandes zu verbessern. Ziel ist es, einen auswertbaren Datenbestand für die spätere Analyse zu generieren (vgl. [16]). Für die Trendanalyse sind die Inhaltsdaten (Content) von besonderer Wichtigkeit. Aus ihnen werden später sämtliche Suchbegriffe und Schlüsselwörter abgeleitet und deren Häufigkeit im Datenbestand analysiert. Die Wort-Datenbanktabelle mit allen vorkommenden Wörtern bildet im Kontext der Trendidentifikation in Texten die Basis für alle, der vorgestellten Methoden. Für Trendanalysen allgemein ist die Zeit der elementare Schlüssel zur Identifikation und Manifestation von Trends. Daraus geht hervor, dass für jedes gespeicherte Wort ein Zeitbezug hinsichtlich seines Vorkommens vorhanden sein muss. Die Grunddaten liegen nach der Selektion und der Aufbereitung zusammengefügt und in einem einheitlichen bereinigten Zustand vor. Durch weitere Reduktion oder Transformation werden die Daten soweit wie möglich vereinfacht und vorbereitet (vgl. [19] & [17]). Um die Menge der Wörter zu reduzieren, wurde entschieden, die Wörter auf ihren Informationsgehalt zu untersuchen, um eine Art Stoppliste zu schaffen, die unwichtige Wörter herausfiltert, auch wenn zukünftig neue Datensätze hinzukommen sollten. Das Hauptelement der Burst-Analyse ist der zu berechnende Vergleichsfaktor, mit dessen Hilfe die Ausbrüche identifiziert werden können. Die Berechnung des Faktors basiert jeweils auf der gewählten Zeitspanne und dem gewählten Suchwort. Daraus ergeben sich durch den erstellten Datensatz über einen Beobachtungszeitraum von einem halben Jahr und der Anzahl der unterschiedlichen Wörter über 12 Millionen Möglichkeiten, diesen Faktor zu bestimmen. Ähnlich viele Möglichkeiten ergeben sich bei der SVD- und der Keyword-Analyse. Bei der SVD-Analyse kommt außerdem die Problematik der hohen Komplexität hinzu. Die Kennzahl IDF innerhalb der Keywordanalyse ist unabhängig von einem für die Berechnung definierten Zeitraum oder einer Aggregation und bleibt somit konstant. Aus diesem Grund wird diese Kennzahl vorberechnet. Nach Abschluss der Datenvorbereitung kommt es zur Anwendung der Data-Mining-Methoden. Die drei Methoden, die im Kontext der definierten Zielstellung der Inhaltsanalyse auf Trendcharakteristiken (Stärke, Form, Zeit, Treiber, Strukturbrüche) ausgewählt wurden und in diesem Kernprozess des KDD-Konzepts Anwendung finden, sind Methoden, die ausschließlich zur

Beschreibung der Daten dienen. Die Burst-Analyse identifiziert je nach festgelegtem Berechnungszeitraum strukturelle Veränderungen der Merkmalsausprägungen im Verlauf. Dagegen gibt die SVD-Analyse den dominierenden und nichtdominierenden Verlauf im zu untersuchenden Datensatz durch Approximation der Datenmatrix wieder und ermittelt jeweils die einflussreichsten Web-Quellen (Feeds). Die letzte der drei Methoden, die Keyword-Analyse, ist in der Lage, Abhängigkeiten zwischen verschiedenen Objekten durch ein Korrelationsmaß (TF-IDF) zu berechnen. Aus den Ergebnissen dieser Analyse können treibende Kräfte ermittelt werden, die entweder den kontinuierlichen Verlauf eines Trends oder die strukturelle Veränderung in der Zeitreihe begründen können. Der letzte Teilschritt im KDD-Konzept von [12] ist die Interpretation und Evaluierung. Zu diesem Teilschritt gehört auch die Visualisierung der entdeckten Muster, um sie für den Nutzer in verständlicher Weise darzustellen (vgl. [19]). Abschließend wird das durch die Interpretation des Anwenders entdeckte Wissen evaluiert und konsolidiert.

4 Ergebnisse und Evaluation

Wie unter 2.3 beschrieben wurde, können nach dem Konzept von Ansoff [1] schwache Signale sich soweit verdichten, dass diese ein Unternehmen zum Handeln bewegen (vgl. [9]). Im Kontext des Web Content Mining von Blogs entstehen solche schwachen Signale in den Veränderungen von Meinungen in Diskussionen (vgl. [23]). Um eine valide Evaluierung der Methoden hinsichtlich der Tauglichkeit zur Unterstützung des CRM mit dem Schwerpunkt der Erkennung von Strukturbrüchen in Trends vornehmen zu können (vgl. [8]), ist eine Datenbasis nötig, die reale Sachverhalte aus verschiedenen Web-Quellen enthält. In diesem Fall wurde eine Miniwelt aus 435 unterschiedlichen Feeds von diversen Blogs (vgl. [13]) geschaffen, die einen Teilausschnitt der Realität abbildet. Dieser abgegrenzte Realitätsausschnitt wurde ein halbes Jahr beobachtet und gewonnene Daten abgespeichert. Mit diesen Daten soll nun überprüft werden, ob sich reale Veränderungen mit den Methoden in den Daten erkennen lassen. Das Ziel der Evaluierung soll sein, von einzelnen Beobachtungen der Untersuchungen, allgemeine Aussagen hinsichtlich der Tauglichkeit der Methoden treffen zu können. Die Burst- oder Ausbruchsanalyse von [3] ermöglicht das Identifizieren von Merkmalsausprägungen, die sich über einem bestimmten Grenzwert befinden. In der nicht-experimentellen Evaluation soll überprüft werden, ob das Verfahren der „Burst-Analyse“ von [3] Veränderungen von realen Sachverhalten in einem Trend erkennen kann. Diese Bewertung wird auf der Basis folgender Hypothese durchgeführt: Das Analyseverfahren erkennt Trendbrüche bzw. Veränderungen von realen Sachverhalten. Das Kriterium der Hypothese ist dann erfüllt, wenn sich eine signifikante Relation zwischen Ausbruch in Bezug auf einen realen Sachverhalt herstellen lässt. Dazu wird eine neutrale externe Web-Quelle herangezogen, die nicht Bestandteil des verwendeten Teilausschnitts der Realität ist. Das Webarchiv der Zeitung „Die Welt“[1] dient dazu als Vergleichsobjekt. Das Kriterium der Evaluationshypothese ist dann erfüllt, wenn sich eine signifikante Relation zwischen Ausbruch in Bezug auf einen realen Sachverhalt herstellen lässt (vgl. dazu jeweils die letzte Spalte in den *Bildern 1 und 3*). Außerdem werden bei der Evaluierung die Stärke der identifizierten Ausbrüche und die durchschnittliche Häufigkeit in Bezug auf den Beobachtungszeitraum erfasst, um die Ergebnisse deutlicher interpretieren zu können (siehe *Bild 1*). Die Trendanalyse durch Singulärwertzerlegung von [10] ist ein Verfahren, das es erlaubt, Trendverläufe und deren Einflüsse zu analysieren. Dabei werden, je nach Analyseziel, Datenmatrizen, bestehend aus den Häufigkeiten, der Zeit und den

[1] http://www.welt.de/nachrichtenarchiv/

Quellen, gebildet, die dann durch das Approximationsverfahren der Singulärwertzerlegung soweit komprimiert werden, dass nur noch zwei Verläufe in Abhängigkeit ihres Einflusses auf die Quellen hergeleitet werden.

Wort	Burst		Realer Sachverhalt (Die Welt (2011))	
	Stärke	Datum		
Mircosoft	3,25 σ	08.10.2010	Adobe und Microsoft gemeinsam gegen Apple	x
	5,43 σ	11.10.2010	Windows Phone 7: neues Smartphone-Betriebssystem vorgestellt // Microsoft hofft im Mobilfunk auf Comeback	x
μ = 21,7	2,53 σ	12.10.2010		
	2,53 σ	13.10.2010		
	4,93 σ	11.02.2011	Microsoft und Nokia bilden eine Allianz	x
	2,86 σ	15.03.2011	Microsoft veröffentlicht Internet Explorer 9	x
	Ø 3,59 σ			100,0%

Bild 1: Relation zwischen Ausbruch und realen Sachverhalt

Diese beiden Verläufe bilden nach [10] den dominierenden Trendverlauf und den nicht-dominierenden Trendverlauf. Es stellt sich bei dieser Evaluation die Ausgangsfrage, ob durch die Komprimierung der Datenmatrix der Verlauf verfälscht wird und dadurch Ausbrüche bzw. Signale verloren gehen. Hinsichtlich dieser Fragestellung wurde folgende Hypothese und deren Kriterium aufgestellt: Das Analyseverfahren erkennt den tatsächlichen Trendverlauf, einschließlich enthaltender Trendbrüche. Da die identifizierten Ausbrüche durch die Burst-Analyse auf den absoluten Häufigkeiten des verwendeten Realitätsausschnitts basieren, werden diese Strukturbrüche und der reale Sachverhalt, sofern er sich über das Zeitungsarchiv zuordnen lässt, als Vergleichsobjekt für die Überprüfung des Kriteriums herangezogen. Um zu überprüfen, ob die Ausbrüche in den beiden Verläufen der Trendvektoren $\overrightarrow{t1}$ und $\overrightarrow{t2}$, ermittelt durch die Singulärwertzerlegung von [10], mit den Ausbrüchen in den absoluten Daten übereinstimmen, wurde die SVD-Analyse um die Berechnung der Burst-Linie in der Applikation erweitert. Der Wert dieser Linie wird analog wie der Faktor in der Burst-Analyse berechnet. Allerdings basiert dieser nicht auf den absoluten Häufigkeiten, sondern auf den Merkmalsausprägungen der beiden Trendvektoren (Siehe *Bild 2*). Im Rahmen der Evaluierung wird jeder durch die Burst-Analyse ermittelte Ausbruch in den Stichproben mit denen in der Singulärwertzerlegung verglichen und Übereinstimmungen werden in einer Tabelle festgehalten (siehe *Bild 3*). Mit dem von [3] modifizierten TF-IDF Verfahren ist es möglich, Schlüsselwörter zu berechnen, die in Korrelation zu einem Suchwort stehen. Im Rahmen der Evaluation kann die Vorgehensweise, die von [3] beschrieben wurde, nicht verwendet werden. Nach ihrem Vorgehen werden nur die letzten 30 aktuellen Dokumente in die Berechnung einbezogen. Unter dem Aspekt der Trendanalyse liefern Schlüsselwörter, die aus den letzten 30 Dokumenten gebildet wurden, keine validen Ergebnisse, um Zusammenhänge zu einem Such- oder Signalwort über einen langen Beobachtungszeitraum herzustellen.

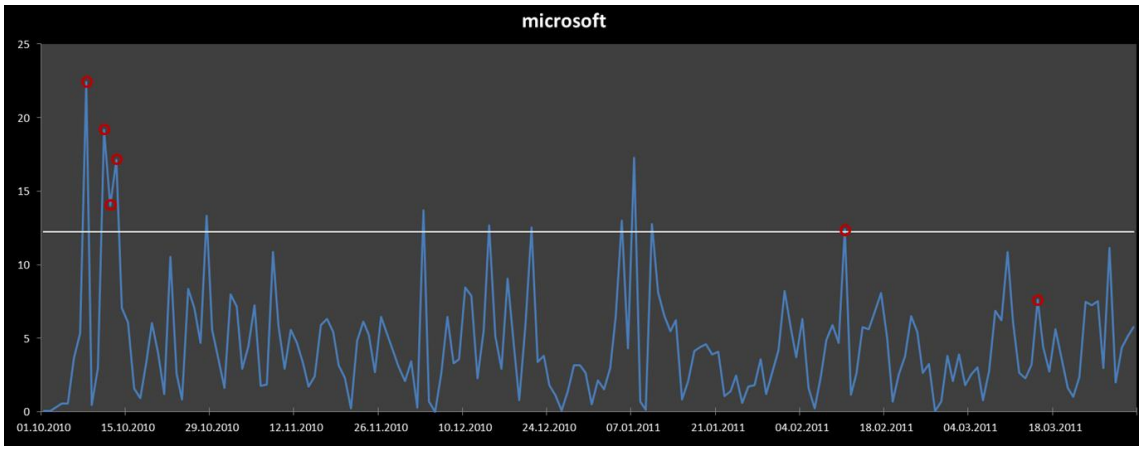

Bild 2: Trendvektor $\overrightarrow{t1}$ für das Suchwort "Microsoft" einschließlich der markierten Ausbrüche aus der Burst-Analyse

Wort	Burst		SVD	
	Stärke	Datum	t1	t2
Mircosoft	3,25 σ	08.10.2010	x	x
	5,43 σ	11.10.2010	x	x
μ = 21,7	2,53 σ	12.10.2010	x	x
	2,53 σ	13.10.2010	x	x
	4,93 σ	11.02.2011	x	o
	2,86 σ	15.03.2011	o	o
	Ø 3,59 σ		83,3%	66,7%

Bild 3: Vergleich der Ausbrüche

Analyseverfahren, die einen Zeitbezug nicht erlauben, sind für die Trendanalyse, in der Zeit eine elementare Rolle spielt, nicht geeignet. Aus diesem Grund wird in der Evaluation keine Eingrenzung durch die Anzahl der Dokumente vorgenommen, sondern ausschließlich durch die Zeit. Es stellt sich bei der Keyword-Analyse die Frage, ob dieses Verfahren tatsächlich valide Schlüsselwörter liefert, um reale Sachverhalte rekonstruieren zu können. Für Unternehmen ist es nicht ausreichend, festzustellen, wann ein Ausbruch im Verlauf vorkommt, sondern warum. In der Theorie sollte diese Methode solche Informationen liefern können, denn im Gegensatz zu den anderen Verfahren in der Evaluation sind die Ergebnisse keine Verläufe oder Ausbrüche, sondern auf den Inhalt bezogene Schlüsselwörter. Um die Aussagekraft dieser Ergebnisse zu überprüfen, wurde für die Evaluation folgende Hypothese aufgestellt: Das Analyseverfahren ermöglicht eine teilweise Rekonstruktion der realen Sachverhalte. Für die Überprüfung der Hypothese werden die ermittelten realen Sachverhalte aus der Burst-Analyse als Vergleichsobjekt herangezogen. Lassen sich Zusammenhänge zwischen den realen Geschehnissen und den berechneten Schlüsselwörtern herstellen, so wird das Kriterium als erfüllt angesehen. Einzelne Wörter können keinen realen Sachverhalt wiedergeben. Deshalb unterliegt die Interpretation der Ergebnisse der subjektiven Wahrnehmung der Person, die das Evaluationsverfahren durchführt. Als Berechnungsparameter wurde in der Durchführung festgelegt, dass nur die 10 höchstgewichtigsten Schlüsselwörter, die in Korrelation zum Suchwort stehen, für den Vergleich genutzt werden. In der untersuchten Stichprobe wurden 23 Ausbrüche mit der Burst-Analyse identifiziert, deren Merkmalsausprägung um das Doppelte der Standardabweichung zum arithmetischen Mittel abweicht. Zu 90% der Ausbrüche konnte ein realer Sachverhalt mit Hilfe des

Zeitungsarchivs „Die Welt“ zugeordnet werden. Die Ausbrüche, die sich über zwei oder drei Tage hintereinander erstreckten und denselben Sachverhalt aufwiesen, wurden zu einem Ausbruch zusammengefasst. Damit wurde eine signifikante Relation zu realen Sachverhalten in Bezug auf die Häufigkeit eines Suchwortes bestätigt. Bei der Evaluierung der SVD-Analyse durch Vergleich der beiden Trendvektoren mit den Strukturbrüchen konnte festgestellt werden, dass lediglich im Durchschnitt 43,5% der Ausbrüche im ersten Trendvektor und 52,4% im zweiten Trendvektor erkannt wurden. Somit konnte keine signifikante Relation zu den zeitlichen Verläufen von realen Sachverhalten in Bezug auf die Häufigkeit eines Suchwortes im Realitätsausschnitt nachgewiesen werden. Die Hypothese konnte somit im Rahmen dieser Evaluation nicht bestätigt werden.

5 Ausblick und Fazit

Das Beobachten von Feeds im Web, was als „Scanning der Umwelt“ zum Erkennen von schwachen Signalen bezeichnet wird, bildet nach dem Konzept Ansoff [1] die Grundlage für das Weak Signal Management & Strategic Issue Management (vgl. [9]). In der Erkennung von Trends in Bezug auf das CRM System spielen dabei Strukturbrüche eine große Rolle. Es stellt sich an dieser Stelle die Frage, inwieweit die verwendeten Methoden dieses Konzept beim Erkennen der schwachen Signale in Form von Strukturbrüchen oder Veränderungen unterstützen können und welche Informationen sich über diese Signale generieren lassen. Abrupte Veränderungen in der Diskussion oder der in der digitalen Mundpropaganda können mit Hilfe der Burst-Analyse identifiziert werden. Wie in der Evaluation bestätigt wurde, kann die Keyword-Analyse ein nützliches Instrument dafür sein, die Ursachen solcher Veränderungen zu ergründen und zu verstehen. Auch wenn die interne Validität der Ergebnisse für das SVD-Verfahren nicht bestätigt werden konnte, ergeben sich dennoch Möglichkeiten, wie diese Methode zur Unterstützung des CRM beitragen könnte. So gibt das SVD-Verfahren Aufschluss darüber, welche Feed-Quellen welchen Einfluss auf den Verlauf haben. Des Weiteren könnten Verfahren wie die Lemmatisierung (vgl. [18]) und das Clustering von Wörtern in der Datentransformation die Analysemöglichkeiten im Konzept deutlich erhöhen. Im Rahmen dieser Arbeit wurden nur drei Verfahren verglichen. Somit kann diese Arbeit nicht den Status einer vollständigen Evaluation aller Verfahren beanspruchen. Zukünftig sollten also weitere Verfahren auf die Eignung zur Erkennung von Trends mit untersucht werden.

6 Literatur

[1] Alby, T. (2008): Web 2.0: Konzepte, Anwendungen, Technologien, 3. Aufl., Hanser Verlag, München.

[2] Ansoff, H. I. (1990): *Implanting strategic management*, 2. Aufl., Prentice Hall Engelwood Cliffs, New York.

[3] Bansal, N.; Koudas, N. (2007a): *Searching the Blogosphere*, Tenth International Workshop on the Web and Databases, WebDB 2007, Beijing, China, June 15, 2007.

[4] Bansal, N.; Koudas, N. (2007b): BlogScope: A System for Online Analysis of High Volume Text Streams, in: Koch, C.; Gehrke, J.; Garofalakis, M.N.; Srivastava, D.; Aberer, K.; Deshpande, A.; Florescu, D.; Chan, C.Y.; Ganti, V.; Kanne, C.C.; Klas, W.; Neuhold, E.J. (Hrsg.): *Proceedings of the 33rd International Conference on Very Large Data Bases*, University of Vienna, Austria, September 23-27, 1410-1413.

[5] Bea, F.X.; Haas, J. (2005): *Strategisches Management*, 4. Aufl., Lucius & Lucius Verlagsgesellschaft mbH, Stuttgart.

[6] Bergold, U.; Mayer, B. (2005): Flow statt Frust: Mit Behavioral Finance und Technische Analyse zu den Gewinnern gehören, 2. Aufl., Finanzbuch Verlag, München.

[7] Bischoff, K.; Firan, C.S.; Kadar, C.; Nejdl, W.; Paiu, R. (2009): Automatically Identfying Tag Types, in: Huang, R.; Yang, Q.; Pei, J.; Gama, J.; Meng, X.; Li, Y. (Hrsg.): *Advanced Data Mining - Applications - 5th International Conference*, ADMA 2009, Beijing China, Springer Verlag; Heidelberg, 31-54.

[8] Bortz, N.; Döring, J. (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*, 4. Aufl., Springer Verlag, Heidelberg.

[9] Buchholz, L. (2009): Strategisches Controlling: Grundlagen – Instrumente – Konzpte, 1. Auf., Gabler Verlag; Wiesbaden.

[10] Chi, Y.; Tseng, B.L.; Tatemura, J. (2006): Eigen-trend: trend analysis in the blogosphere based on singular value decompositions, in: Yu, P.S.; Tsotras, V.J; Fox, E.A.; Liu, B. (Hrsg.): *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, November 6-11, 68-77.

[11] Deerwester, S.; Dumais, S.; Landauer, T.; Furnas, G.; Harshman, R. (1990): Indexing by latent semantic analysis, in: *Journal of the American Society for Information Science (JASIS)*, Volume 41, Number 6, September 1990, 391-407.

[12] Fayyad, U. M.; Piatetsky-Shapiro, G.; Symth, P. (1996): From Data Mining to Knowledge Discovery: An Overview, in: Fayyad, U. M.; Piatetsky-Shapiro, G.; Symth, P.; Uthurusamy, R. (1996): *Advances in Kowledge Discovery and Data Mining*; AAAI Press, Menlo Park, California, 1-34.

[13] freshfeeds.de (2010): *Das RSS-Feeds Verzeichnis – Freshfeeds*, URL: http://www.freshfeeds.de, Abruf am 25.09.2010.

[14] Freitag, K. (2003): *Zeitreihenanalyse: Methoden und Verfahren*, 1. Aufl., EUL Verlag, Köln.

[15] Fung, G.P.C.; Yu, J.X.; Yu, P.S.; Lu, H. (2005): Parameter free bursty events detection in text streams, in: Böhm, K.; Jensen, C.S.; Haas, L.M.; Kersten, M.L.; Larson, P.A.; Ooi, B.C. (Hrsg.): *Proceedings of the 31st International Conference on Very Large Data Bases,* Trondheim, Norway, August 30 - September 2, 181-192.

[16] Gabriel, R; Gluchowski, P.; Pastwa, A. (2009): *Data Warehouse & Data Mining*, 1. Aufl., W3l Verlag, Witten.

[17] Hagemann, S. (2005): Maßzahlen für die Assoziationsanalyse im Data Mining: Fundierung, Analyse und Test, 1. Aufl., Diplomica Verlag, Hamburg.

[18] Hausser, R. (2000): Grundlagen der Computerlinguistik: Mensch-Maschine-Kommunikation in natürlicher Sprache, 1. Aufl., Springer Verlag, Heidelberg.

[19] Holthuis, J. (2001): *Der Aufbau von Data Warehouse-Systemen: Konzeption – Datenmodellierung – Vorgehen*, 2. Aufl., Gabler & Deutscher Universitätsverlag, Wiesbaden.

[20] Kanjilal, P.P. (1995): Adaptive prediction and predictive control, 1. Aufl., Peter Peregrinus Verlag, London.

[21] Klahold, A. (2009): Empfehlungssysteme: Recommender Systeme - Grundlagen, Konzepte und Lösungen, 1. Aufl., GWV Fachverlag GmbH, Wiesbaden.

[22] Kleinberg, J. M. (2003): Bursty and hierarchical structure in streams, in: *Data Mining and Knowledge Discovery*, Volume 7, Number 4, October 2003, Springer Verlag, 373-397.

[23] Koller, P.; Alper, P. (2008): Die Bedeutung privater Weblogs für das Issue-Managment in Unternehmen, in: Alper P.; Blaschke S. (Hrsg) : *Web 2.0 – Eine empirische Bestandsaufnahme*; 1. Aufl., Vieweg & Teubner Verlag, Wiesbaden.

[24] Petersohn, H. (2005): *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur*, 1. Aufl., Oldenbourg Wissenschaftsverlag GmbH, München.

[25] Pillkahn, U. (2007): *Trends und Szenarien als Werkzeuge zur Strategieentwicklung*, 1. Aufl., GWA Kommunikation GmbH Verlag, Erlangen.

[26] Platakis, M.; Kotsakos, D.; Gunopulos, D. (2009): Searching for Events in the Blogosphere Manolis Platakis, In: Quemada, J.; León, G.; Maarek, Y.S.; Nejdl, W. (Hrsg.): *Proceedings of the 18th International Conference on World Wide Web*, WWW 2009, Madrid, Spain, April 20-24, 1225-1226.

[27] Proximity (2005): Proximity Studie: Corporate Blogging, URL: http://www.pressrelations.de/new/standard/result_main.cfm?r=190069&sid=&aktion=jour_pm&poffset=4398660000190069&quelle=0, Abruf am 15.04.2011.

[28] Qu, C.; Li, Y.; Zhu, J.; Hunag, P.; Yuan, R.; Hu, T. (2008): Term Weighting Evaluation in Bipartite Partitioning for Text Clustering, in: Li, H.; Liu, T.; Ma, W.Y.; Sakai, T.; Wong, K.F.; Zhou, G. (Hrsg.) (2008): *Information Retrieval Technology*; 4th Asia Information Retrieval Symposium, AIRS 2008 Harbin, China, January 2008, Springer Verlag, Heidelberg, 393-400.

[29] Sieben, F. (2002): *Rückgewinnung verlorener Kunden*, 1. Aufl, Deutscher Universitätsverlag, Wiesbaden.

[30] Schneider, J.; Minnig, C.; Freiburghaus, M. (2007): *Strategische Führung von Nonprofit-Organisationen*, 1. Aufl., UTB Verlag, Stuttgart.

[31] Stahel, W. (2008): Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler, 5.Aufl., GWV Fachverlag GmbH, Wiesbaden.

[32] Technorati (2009): State of the blogosphere 2009, URL: http://technorati.com/blogging/feature/state-of-the-blogosphere-2009, Abruf am 27.10.2010.

[33] Watkins, D. (2002): Fundamentals of matrix computations, 2. Aufl., John Wiley & Sons Inc. New John Wiley & Sons Inc, New York.